Measurement error: estimation, correction, and analysis of implications

Pankowska, P.K.

2020

Summary

Despite countless attempts to reduce it and address its causes, measurement error is a problem inherent to all data sources. (Alwin, 2007; Biemer et al., 1991; Kuha & Skinner, 1997). Its presence often leads to biased and inconsistent statistical estimates and, as a consequence, to erroneous findings and conclusions. It is therefore crucial to understand, account, and correct for measurement error to ensure research validity (Fuller, 2009; Grace, 2017; Kuha & Skinner, 1997).

Latent variable modelling is an increasingly popular solution to this problem, as it allows for the estimation of and correction for measurement error without the need for gold standard data. That is, the main advantage of latent variable models (LVMs) is the fact that, unlike alternative measurement-error-correction techniques, they do not make use of error-free validation data, which are rarely available in practice. Instead these models use repeated indicators of the same variable, either cross-sectionally from various sources or over time from the same source, to extract information about measurement error directly from the data (Biemer & Bushery, 2000).

A group of LVMs that are applied specifically to categorical, longitudinal data, and which are the main focus of this dissertation, are hidden Markov models (HMMs) (Biemer, 2004, 2011; Oberski et al., 2017; Pavlopoulos & Vermunt, 2015). HMMs are used when a (dynamic) quantity of interest is measured in a panel survey with some degree of error. The models allow for the separation of true change from measurement error which, in turn, can produce error-corrected estimates of the quantity of interest, and are also able to assess the level of measurement error in the corresponding variable (Biemer, 2011; Pankowska et al., 2018).

The standard HMM consists of two components: (i) the structural component that models the true (latent) initial state probabilities and the true (latent) transition probabilities; and (ii) the measurement component that models the interactions of the observed values (which contain error) with the true values at each time point. The two components are estimated simultaneously. The model relies on two basic assumptions: first, the probability of a specific value occurring at time t only depends on its value in the previous time point – the so-called *Markov assumption*. Second, the probability of observing a specific value at time t only depends on the true value at the same time point – the so-called *local independence assumption* or the *independent classification error (ICE) assumption*. While the standard, single-indicator HMM relies on the local independence assumption for identifiability, this assumption is often viewed as highly restrictive and unrealistic, as it does not allow for the modelling of the presence of systematic errors without risking poor model identifiability. To overcome this challenge, it is possible to use extended, multiple-indicator versions of HMMs. However, this solution introduces some new challenges. Most importantly, the use of multiple indicators usually requires performing record linkage, which might lead to linkage error – a new potential source of bias. Furthermore, the implementation of such extended models also tends to be complex and time-consuming.

Given the potentially strong, adverse effects of measurement error and the possibility of minimizing these using HMMs, the aim of this thesis is twofold: first to understand in more detail the problem of measurement error and second to investigate whether extended HMMs that are applied to linked data can be used for error, and to what extent this method can be feasibly implemented.

In more detail, **Chapter 2** examines the bias introduced by measurement error, using clustering as an illustrative example. More specifically, the simulation study investigates the sensitivity of two commonly used model- and density-based clustering algorithms (i.e. GMMs and DBSCAN) to varying severities and magnitudes of random and systematic errors. The results confirm that measurement error in many cases leads to non-negligible bias, as the returned clusters are (highly) dissimilar to the ones obtained when the dataset is error-free. The number of clusters found in the data is also affected by the error.

Chapter 3 looks at how different data collection processes might impact the nature and magnitude of measurement error, by studying how the switch from dependent interviewing (DI) to independent interviewing (INDI) in the Dutch Labour Force Survey (LFS) affects the random and systematic components of the error. The results indicate that the use of DI lowers the probability of obtaining random errors but has no significant effect on systematic errors. What is more, the results also show that regardless of the interviewing regime used, the survey data, similarly to the register data, also contains autocorrelated error. The findings of this paper indicate that both data sources examined are subject to non-negligible systematic error that needs to be considered when correcting for measurement error using HMMs. This in turn confirms the need for using extended, multiple-indicator HMMs, which allow for the relaxation of the local independence assumption and the modelling of error autocorrelation without risking poor model identifiability.

Chapter 4 investigates whether and to what extent the use of multiple-indicator HMMs, which often requires record linkage, leads to biased estimates due to the presence of linkage error. The results of the simulation study show that overall the sensitivity of the HMM (structural) parameter estimates to false-positive and false-negative linkage error is low. It appears that only rather extreme scenarios (i.e. high error rate and high correlation between the probability of error and model estimates) lead to substantial bias. Moreover, the results also show that under certain conditions, false-positive linkage error acts as another source of measurement error that is absorbed into the error-rate parameters of the model, leaving the latent transition estimates unaffected. In these cases, HMMs also accounts for linkage error.

Finally, **Chapter 5** focuses on a more practical matter. Given their complex nature and, with that, the time and costs associated with their implementation, the study explores the feasibility of using multiple-indicator HMMs in the first instance. More specifically, the study investigates whether it is possible to simplify the error correction procedure by running the full analysis once and then re-using the resultant error parameters as a correction factor for a number of years. The proposed solution is

contingent on the assumption that the structure and size of the error remain constant. The analysis provides some evidence that in the absence of a major change in the data collection process, the size and structure of the error are time-invariant and, therefore, the error parameters can be carried forward a certain number of years.

While the findings presented in this dissertation suggest that HMMs are a promising tool to correct for measurement error in categorical, longitudinal data, several additional aspects need to be considered before this approach can be applied in practice. Namely, the performance and feasibility of the method should be tested in a different context that goes beyond the topic of labor mobility and on data from different countries than the Netherlands. Also, if possible, additional sources apart from surveys and administrative registers should be considered. Furthermore, a thorough examination of model robustness and the sensitivity of parameter estimates to varying model specifications containing different assumptions ought to be carried out. Finally, it is also important to consider how researchers can use error-corrected microdata in their analyses, while accounting for the uncertainty of the "true state" membership.